Technical Report 1250

The Leader AZIMUTH Check: Factor Structure of Common Competencies

John P. Steele
Kansas State University
Consortium Research Fellows Program

Sena Garven
U.S. Army Research Institute

June 2009

United States Army Research Institute
for the Behavioral and Social Sciences

U.S. Army Research Institute
for the Behavioral and Social Sciences

A Directorate of the Department of the Army
Deputy Chief of Staff, G1

Authorized and approved for distribution:

MICHELLE SAMS, Ph.D.
Director

Technical review by

Allison Dyrlund, U.S. Army Research Institute
Gregory Ruark, U.S. Army Research Institute

NOTICES

DISTRIBUTION: Primary distribution of this Technical Report has been made by ARI. Please address correspondence concerning distribution of reports to: U.S. Army Research Institute for the Behavioral and Social Sciences, Attn: DAPC-ARI-ZXM, 2511 Jefferson Davis Highway, Arlington, Virginia 22202-3926.

FINAL DISPOSITION: This Technical Report may be destroyed when it is no longer needed. Please do not return it to the U.S. Army Research Institute for the Behavioral and Social Sciences.

NOTE: The findings in this Research Note are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

REPORT DOCUMENTATION PAGE

1. REPORT DATE (dd-mm-yy)	2. REPORT TYPE	3. DATES COVERED (from. . . to)
June 2009	Final	March2008-November2008

4. TITLE AND SUBTITLE	5a. CONTRACT OR GRANT NUMBER
The Leader AZIMUTH Check: Factor Structure of Common Competencies	5b. PROGRAM ELEMENT NUMBER 622785

6. AUTHOR(S)	5c. PROJECT NUMBER
John P. Steele (Kansas State University) Sena Garven (U.S. Army Research Institute)	A790
	5d. TASK NUMBER 333
	5e. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	8. PERFORMING ORGANIZATION REPORT NUMBER
ARI-FLRU 851 McClellan Ave Ft. Leavenworth, KS 66027	

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. MONITOR ACRONYM
U. S. Army Research Institute for the Behavioral & Social Sciences 2511 Jefferson Davis Highway Arlington, VA 22202-3956	ARI
	11. MONITOR REPORT NUMBER Technical Report 1250

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

Subject Matter POC: Sena Garven

14. ABSTRACT *(Maximum 200 words)*:

Enhancing the leadership skills of Soldiers is of primary importance to the U. S. Army. A critical step in the process of leader development is self-awareness through self-assessment. Such insight is important because identifying and assessing trainable competencies that facilitate maximum leadership effectiveness creates a strategic advantage. This report describes the psychometric properties and common competencies assessed by the Leader AZIMUTH Check, a 360-degree feedback instrument for Army leaders. The AZIMUTH was designed and implemented by the Army Research Institute (ARI) to improve leader common competency development, leader-directed feedback, and enhance leader self-awareness. The purposes of the present research project were to establish a factor structure of common competencies, the minimum number of raters required for adequate reliability, conceptual agreement across rating sources, rating patterns and behaviors, and validity evidence of the AZIMUTH.

15. SUBJECT TERMS

Multisource feedback; leadership skills; rating psychometrics; leader development

SECURITY CLASSIFICATION OF			19. LIMITATION OF ABSTRACT	20. NUMBER OF PAGES	21. RESPONSIBLE PERSON
16. REPORT	17. ABSTRACT	18. THIS PAGE			
Unclassified	Unclassified	Unclassified	Unlimited	45	Ellen Kinzer Technical Publication Specialist 703/602-8049

Technical Report 1250

The Leader AZIMUTH Check: Factor Structure of Common Competencies

John P. Steele
Kansas State University
Consortium Research Fellows Program

Sena Garven
U.S. Army Research Institute

ARI-Fort Leavenworth Research Unit
Stanley M. Halpin, Chief

U.S. Army Research Institute for the Behavioral and Social Sciences
2511 Jefferson Davis Highway, Arlington, Virginia 22202-3926

June 2009

Army Project Number
622785A790

Personnel Performance
and Training Technology

ACKNOWLEDGEMENTS

The authors would like to thank Dr. Ronald Downey of Kansas State University for his conceptual and technical contributions to the work described in this report. Many have contributed to the AZIMUTH development over the years, and while it is not possible to name all contributors, we want to thank those whose prior efforts made this work possible.

THE LEADER AZIMUTH CHECK: FACTOR STRUCTURE OF COMMON COMPETENCIES

EXECUTIVE SUMMARY

Research Requirement:

Three hundred sixty degree feedback refers to multi-source feedback used for developmental purposes that originates from superiors, subordinates, peers, and self raters. The main tenet behind the utility of the 360-degree approach is that all leaders have blind spots, or unnoticed skill deficiencies and strengths and feedback from multiple sources can facilitate leader development. In order to improve feedback and enhance self-awareness on Army officer's leadership competence, the Army Research Institute (ARI) developed a 360-degree feedback process for Army officers. The AZIMUTH has been utilized for over a decade, starting with a 1996 pilot survey, resulting in a database of nearly 6,000 ratings. Various analyses of this data have been undertaken but analyses to date have not included a quantitative validation. However, consistent with standard content validation processes, content validity has been logically investigated and established.

Procedure:

A review of archival AZIMUTH ratings from three Army samples was conducted to identify the common competencies assessed by the AZIMUTH. After establishing a factor structure, the psychometric properties and validity evidence of the AZIMUTH was investigated. Quantitative findings were compared to organizational, psychometric, and military literatures.

Findings:

Two factors, task and interpersonal competency emerged from the original thirteen conceptualized AZIMUTH dimensions. Although the two factors were highly correlated, they both uniquely contributed to predicting single-item assessment of leadership effectiveness. Inconsistent with general organizational literature, there was no differential reliability across rating sources. Only two raters were required to produce convergent ratings and there was support for conceptual equivalency across rating sources. Inconsistent with general organizational findings, but consistent with previous military analyses, the data indicated a modesty bias, in which self-ratings were generally lower than ratings from other sources. Finally, absolute agreement of self-other ratings occurred frequently (80 %), indicating a high degree of convergence between sources. It was determined that using the AZIMUTH, officer ratings were valid; however, there was a lack of discrimination between leadership competencies.

Utilization and Dissemination of Findings:

This analysis answered some of the essential unanswered questions regarding the psychometric qualities of the AZIMUTH. Knowing that all sources add reliable ratings argues for

the continuation of all rating sources in garnering developmental feedback information. Similarly, the presence of a modesty bias provides context for subsequent analyses. Most importantly, the findings suggest that the AZIMUTH is a valid one-size-fits all assessment useful for Army officer development. The methods and findings presented can be used as a guide to 360-degree feedback validation procedures and for self-development of important common competencies.

THE LEADER AZIMUTH CHECK: FACTOR STRUCTURE OF COMMON COMPETENCIES

CONTENTS

LIST OF TABLES

APPENDICES

Introduction

Background

Enhancing leadership skills of military leaders is of primary importance to the Army. A previous version of the U.S. Army's official vision statement states, "We are about leadership; it is our stock in trade, and it is what makes us different" (U.S. Army, 1999b, p. 7). More recently, the General TRADOC Vision includes the following:

> "To shape both today's Army and the Future Combat Force, the Army… Develops adaptive leaders: TRADOC trains leaders for certainty and educates them for uncertainty. Leader development produces innovative, flexible, culturally astute professionals expert in the art and science of the profession of arms and able to quickly adapt to the wide-ranging conditions of full spectrum operations." (U.S. Army, 2008a)

Finally, FM 3-0 (2008b) Operations, states that, "Success in battle depends on …competent leadership." Developing an understanding of Army leadership is important because it provides assessable and trainable competencies that facilitate maximum leadership effectiveness (Hatfield, 1997). Self development of leadership skills is a critical step in the process of developing Army leaders (Flowers, 2004). Flowers argued that developing leadership skills is so important that it should be incorporated at all levels of military leadership. Civilian literature also stresses the importance of identification of important leadership skills, as evidenced by many books, articles, and a special issue in *The Leadership Quarterly* (Mumford, Marks, Connelly, Zaccaro & Reiter-Palmon, 2000). Insight into common competencies that impact leader performance is valuable for furthering organizational goals, leader development, and selection (Connelly et al., 2000; Wong, Gerras, Kidd, Pricone, & Swengros, 2003).

Military leadership is stressed and developed through several mechanisms including formal education, operational assignments, and self-development (Wong, Bliese, McGurk, 2003). A necessary component of self-development and mastery is a true understanding of one's own strengths and limitations, and the ability to adapt to situational demands. From this perspective it is obvious that leadership skills must be assessed before they can be developed. Despite decades of research, there is still a need to clearly define, assess, and subsequently develop military leadership (Wong, Bliese, et al., 2003). Multi-source feedback has proven useful in some contexts as an assessment approach within a leader development process. The purpose of the present paper was to investigate the factor structure and validity of the multi-source Leader AZIMUTH Check (AZIMUTH; Halpin, 1997) in an effort to expand application and theory regarding multi-source feedback in the Army.

Three hundred sixty degree feedback refers to multi-source feedback used for developmental purposes that originates from superiors, subordinates, peers, and self raters. According to Karrasch (2006, p.1), "The goal is to provide unbiased, objective feedback from multiple perspectives so that the leaders can gain the personal insight needed to maintain leadership strengths and address leadership developmental needs in order to perform better." The main tenet behind the utility of the 360-degree approach is that all leaders have blind spots, or unnoticed skill deficiencies and strengths (Karrasch, Halpin, & Keene, 1997) and feedback

from multiple sources can facilitate leader development (London & Smither, 1995). In order to improve feedback and enhance self-awareness on Army officer's leadership competence, the Army Research Institute (ARI) developed a 360-degree feedback process for Army officers. In addition to military consideration, 360-degree feedback has been widely used in corporations (Edwards & Ewen, 1998).

Newer leaders especially often lack self-awareness and the 360-degree process can be a useful mechanism for enhancing self-awareness, and subsequently improving leader performance, strengthening leader-other relationships, and reinforcing organizational culture (Bunker, Kram & Ting, 2002; Garavan, Morley & Flynn, 1997). While superiors are a common rating source in traditional feedback systems, their view of the ratee is generally narrow. Ratees can also affect superior ratings by engaging in impression management (i.e., representing themselves more favorably) thus, possibly distorting evaluations (Van der Heijden & Nijhof, 2004). Although superiors arguably provide the most reliable ratings (Greguras & Robie, 1998), they often have a limited view of how the leader interacts with others. It would be expected that both peers and subordinates are in a better position to make evaluations regarding their interactions with the leader (Karrasch & Halpin, 1999). Another benefit to having multiple raters in the feedback system is that this will instill confidence in the developmental process (Garavan et al., 1997). It is much easier for the ratee to consider that a single rater is biased or inaccurate, but it is much more difficult to be dismissive when there is recurring information provided by multiple raters from each source. In sum, all rating sources are valued and a degree of discrepancy between sources is to be expected because each source has a unique opportunity to observe the ratee's performance behaviors (London & Beatty, 1993). One Army 360-degree feedback program is the Leader AZIMUTH Check (Halpin, 1997).

AZIMUTH *Background*

The AZIMUTH (Halpin, 1997) is a generic instrument designed to assess leadership by garnering feedback from self, peer, subordinate, and superior raters (i.e., 360-degree feedback). Army leadership training, like general Army training, is premised on the notion that feedback is essential to learning and self-development (Halpin, 1997). Therefore, the purpose of Army 360-degree feedback is to gain insights on one's own behavior from others who provide valid ratings (Karrasch et al., 1997). The Leader AZIMUTH Check has been used for over a decade to solicit and organize 360-degree feedback to thousands of Army leaders on doctrine-based competencies (Halpin, 1997). AZIMUTH feedback is provided from the vantage point of those who work closely with the leader and indicates the extent to which leaders are perceived as having characteristics congruent with military leadership and the Warrior Ethos.

Conceptually, the AZIMUTH has nine first order *competencies* (communicating, decision-making, motivating, developing, building, learning, planning and organizing, executing and assessing), three *values* (respect, selfless-service, and integrity), and one *attribute* (emotional stability), all of which were derived from Army Doctrine (FM 22-100, 1999) on leadership (Laffitte & Metcalf, 2006). Subsequent versions of the AZIMUTH (such as Form 6106) have dropped the attribute, but still contain the original competencies and values. The AZIMUTH has been utilized for over a decade, starting with a 1996 pilot survey, resulting in a database of nearly 6,000 ratings. Various analyses of this data have been undertaken (e.g., Lafitte &Metcalf)

but analyses to date have not included a quantitative validation (Horey, Harvey, Curtin, Keller-Glaze, Morath, & Fallesen, 2006). However, consistent with standard content validation processes (AERA/APA/NCME, 1999; SIOP, 2003), content validity has been logically investigated and established.

Content Validation

The Leader AZIMUTH Check evolved from the Strategic Leader Development Inventory (SLDI; Owens, 1996), developed jointly by the Army Research Institute (ARI), the Army War College (AWC) and the Industrial College of the Armed Forces (ICAF) in a long-term project directed by Dr. T. Owen Jacobs at Army Research Institute (ARI). SLDI item development began with a literature search, a comparison to military instruction on leadership and leader development, and consultation with subject matter experts. The SLDI was based in part on extensive interviews with more than one hundred general officers and input from students at the Army War College (Halpin, 1997). Experts established content validity by judging the construct of Army leadership, operationalized in the Army Doctrine on leadership, and the SLDI items. The SLDI, as the name suggests, was designed to provide feedback to relatively senior officers (lieutenant colonels and colonels with 18+ years of experience) who were students at the AWC and ICAF and who were preparing for a possible new career phase involving strategic rather than tactical or operational skills and knowledge. A former student from ICAF, who had been exposed to the SLDI, solicited ARI assistance in implementing the instrument for use in providing feedback to students (Army captains) at the Combined Arms Staff and Services School (CAS^3). Analysis of preliminary data from SLDI used within CAS^3 made it clear that the "strategic" nature of the instrument resulted in a mismatch with the experience level of the CAS^3 target population.

Approximately 50 of the original items were extracted from the SLDI, based on item analyses of the preliminary data from several hundred captains in CAS3; these items were supplemented with an additional 46 items written to better reflect the leadership understanding and experience of captains with 4-6 years of experience. Although there were several clusters of items which reflected common leadership themes, (e.g., communication), the intent was to present a broad set of items and subsequently derive a post hoc factor analytic set of dimensions, which would then be used to structure feedback to the CAS^3 students. This approach was roughly parallel to the approach taken in the development of the SLDI. The new instrument was named the Leader AZIMUTH Check, version 1.0. After the first major launch of AZIMUTH in the winter of 1996, it became clear that this approach would not work. Extensive analyses of data from three CAS^3 classes (approximately 2000 individual self-ratings and 6000 peer ratings) revealed no underlying stable factor structure. It was concluded that the lack of a consistent understanding of leadership among this relatively inexperienced population rendered the empirical factor analytic approach ineffective in this situation (S. Halpin, 2008, personal communication, September 2, 2008). Since the intent was not to explore latent concepts of leadership, but rather to provide a tool to provide useful feedback to Army leaders, it was decided that the AZIMUTH instrument and associated feedback would be structured around the dimensions of leadership as identified in Army doctrine which was then being rewritten (U.S. Army, FM22-100, 1999).

The AZIMUTH was retooled and version 2.0 was tested in the spring of 1997 (Halpin, 1997). A second form, version 2.1, was subsequently developed; the only difference was that the generic language of 2.0 was replaced with wording which reflected the use of AZIMUTH in the classroom. The separate item clusters (Communication, Decision Making, etc.) in Versions 2.0 and 2.1 were found to have acceptable Chronbach alpha values of .85 or higher (S. Halpin, 2008, personal communication, September 2, 2008). Although the primary target population of interest was always Army captains in CAS^3, the AZIMUTH was aperiodically used to provide feedback within a variety of Army units and organizations; on these occasions version 2.0 was supplemented by additional items reflecting additional elements of leadership relevant to the particular organization. For example, in an application conducted within a multi-national headquarters, items were added to reflect cross-cultural communication. The AZIMUTH **2.0** remains as a generic (generic meaning to be used by Army personnel of various ranks, positions, and background) assessment of Army leadership.

The AZIMUTH has been evaluated several times in over a decade in both published and unpublished reviews (e.g., Halpin 1997; Karrasch & Halpin 1999; Karrasch, et al., 1997; Laffitte, Halpin, & Tran, n.d.; Laffitte & Metcalf, 2006). During each administration of the AZIMUTH, in addition to general psychometric testing, respondents indicated the degree to which the items: 1) represented observable behaviors, 2) were clear and understandable, 3) measured elements critical to excellence in Army leadership, and 4) were related to global assessments of effectiveness, good leadership, and maintenance of effective interpersonal relations. All findings have indicated that the AZIMUTH contains observable and face valid items. Karrasch and Halpin (1999) have also found support for user acceptance and satisfaction, and a level of comfort of concerning the confidentiality in data collection.

Despite a decade of research and widespread use, there has been little investigation of the underlying factor structure, inter-rater reliability, construct agreement, or construct validity of the AZIMUTH. In other words, we still do not know:

-how many unique dimensions are assessed,

-how many raters are required for consistent (reliable) ratings,

-if raters from each source (i.e., self, subordinate, superior, and peer) use the same mental model when making evaluations,

- if the AZIMUTH is valid in the sense that inferences drawn from its use reflect actual changeable behaviors that are related to effective leadership.

Thus, the purpose of the present project is to gain further knowledge regarding:

-a factor structure of common competencies

-the minimum number of raters required for adequate reliability

-conceptual agreement across rating sources

-rating patterns and behaviors

Factor Structure

As stated earlier, a goal of this investigation was to develop a framework of common competencies of effective U.S. Army mid-level Officers. Common competencies are first-order knowledge and skills that are necessary for effective leadership (Brownell, 2006). These

competencies do not allow for complete training that results in optimal effectiveness because they lack adaptability and awareness components; however, common competencies must first be established and trained before the higher-order, or distinctive competencies can be developed (Brownell, 2006). In other words, before knowing which skills to use in a given situation, one must first possess a library of skills. The development of common leadership competencies has received inadequate attention (Connelly et al., 2000; Kanungo & Misra, 1992; Wong, Gerras, et al., 2003;Wright, 1996; Wright & Taylor, 1985; Yukl & Van Fleet, 1992).

In December 2001, the Chief of Staff of the Army tasked the U.S. Army War College with identifying important leader skills for officers (Wong, Gerras, et al., 2003). The research team involved in this task focused on strategic leadership, which they defined as applying to all Army leadership positions. Wong, Gerras, and colleagues argued that the Strategic Leadership Primer and the FM 22-100 Army leadership doctrinal manual were useful, but difficult to apply. Application was limited because "The list is extremely comprehensive and appears to capture every possible aspect of leadership" (Wong, Gerras, et al., 2003, p. 3) and "At the individual level, it is difficult to assess one's leadership ability when the lists suggest that a strategic leader must be, know, and do just about everything" (p. 5). Wong, Gerras, et al. reviewed popular management books and military doctrine and from this proposed six meta-components of strategic leadership: identity, mental agility, cross-cultural savvy, interpersonal maturity, world-class warrior, and professional astuteness. Unlike the Wong, Gerras, et al. focus on the higher-order, or distinctive competencies, , the present investigation focused on the lowest-order, or common competencies. Additionally, the Wong, Gerras et al. review was purely based on popular business literature and Army doctrine, whereas the present efforts were driven by peer-reviewed journals and empirical findings from Army officer developmental feedback ratings. Validating the common competency model is a necessary and important step to understand assessment and develop the higher-order strategic leadership competencies described by Wong, Gerras, et al. Subsequently, Horey, et al. (2004) conducted a more formal competency analysis and their work led to the revised competency descriptions provided in the new Army leadership doctrine (FM 6-22, 2006).

In order to develop common competencies of effective leadership it is important to operationalize leadership. Unfortunately, according to some this is virtually impossible. Stogdill (1974) went so far as to write, "There are almost as many definitions of leadership as there are persons who have attempted to define the concept" (p. 259). However, the literature does indicate commonalities amongst the various operationalizations of leadership. Yukl (2006) concluded that most definitions encompass facilitative activities, relationship-oriented behaviors, structuring behaviors, and the intentional influence of others towards goals. The Ohio State and University of Michigan 1950s studies of leadership were the beginning of research involving leader characteristics, behaviors, skills, and competencies. The purpose of that work was to categorize relevant leadership behaviors and create an assessment of those behaviors. A factor-analysis of civilian and military survey responses revealed two broad categories of leadership that were labeled *consideration* and *initiating structure* (Fleishman, 1953). Researchers have created various conceptualizations of leadership skills or characteristics, with most focusing on two to four broad attributes because of Fleishman's taxonomy (Bowers & Seashore, 1966; Connelly et al., 2000; Mann, 1965; Mumford, Campion, & Moregeson, 2007; Mumford, et al., 2000; Swiderski, 1987).

Hypothesis 1: A two to four factor solution incorporating the original Ohio State model will fit the data.

Inter-rater Reliability

Reliability is a necessary, but insufficient condition for trusting that the feedback is valid enough to share with the ratee, conduct research, or drive policies (Guion, 1997). High inter-rater reliability of ratings between rating sources; however, is generally not expected (Bozeman, 1997; Greguras & Robie, 1998). After all, a fundamental tenet of the 360-degree process is that each rating source is important because it provides another piece of unique information (Hazucha, Hezlet, & Schneider, 1993).

According to Greguras and Robie (1998), what matters most for ratings is not what is rated, but who is rating, and the target leader that is being evaluated. Greguras and Robie also developed a basic rule of thumb for the necessary number of raters for each ratee, when there are 5 items. In other words, how many people does it take to produce consistent ratings? Greguras and Robie found that for a 5-item scale (which closely reflects the subscale size of most of the conceptual AZIMUTH dimensions), 4 supervisors, 8 peers, and 9 subordinates were required to achieve acceptable levels of reliability. This rule of thumb also suggests that superior inter-rater reliability is greater than peer and subordinate inter-rater reliability.

Hypothesis 2: The number of raters required to produce consistent ratings will vary by source.

Conceptual Agreement

In addition to establishing inter-rater reliability, it is also important to establish conceptual agreement across rating sources. That is, the instrument should be perceived the same way regardless of the rater source. Conceptual agreement exists when different sources (i.e., peers, subordinates, and superiors) use the same items with the same importance (i.e., loadings) to represent the same dimension (Cheung, 1999). Failure of conceptual agreement across rating sources complicates between-group comparisons including the standard feedback provided.

This does not mean that different rating sources cannot provide a unique perspective on the performance of an individual. Basic experiences tell us that people act differently around their superior than they act around their subordinates. However, the rating instrument should provide conceptual agreement for all rating sources, because failure to do so would result in feedback ratings that may be inaccurate and misleading (Cheung, 1999; Woehr, Sheehan, & Bennett, 2005).

Hypothesis 3: Army Soldiers will provide conceptually equivalent ratings.

Estimation and Self-Awareness

The purpose of Army 360-degree feedback is to gain insights on one's own behavior from others who provide valid ratings (Karrasch et al., 1997). Therefore, gaining a better understanding of rating behaviors in the Army 360-degree feedback is essential, given the importance and wide application of the AZIMUTH. In addition to adding value to individual feedback, estimation patterns are of further interest. Research has shown that accurate self-perceivers (i.e., self-aware leaders) make more effective job relevant decisions as opposed to under- or over-estimators (Bass & Yammarino, 1991). Likewise, Atwater, Roush, and Fischthal (1995) have argued that accurate self-perception, is by itself a valid predictor of actual performance. More specific to the military context, research conducted with the United States Air Force (Halverson, Tonidandel, Barlow & Dipboye, 2002) revealed that self-subordinate agreement on leadership ratings was a better predictor of promotion rates than self-superior or self-peer agreement, and added incremental validity beyond that provided by self ratings alone. This work highlights the importance of self-awareness, defined as consistency with subordinate perceptions.

Past research has demonstrated a strong tendency for self-ratings to be higher than other ratings from other sources, otherwise described as over-estimation (Fox & Dinur, 1988; Harris & Schaubroeck, 1988; Landy & Farr, 1980, London & Wholers, 1989; Mabe & West, 1982; Podsakoff & Organ, 1986; Thornton, 1980). Finally, Mersman and Donaldson (2000) demonstrated that convergence increased as the domain being evaluated moved from subjective contextual performance to more overtly noticeable task performance, and finally to verifiable task performance.

Hypothesis 4: Self-ratings will be higher than corresponding ratings from the other rating sources.

Method

Participants

In this project we examined archival data that came from three sources, one large TRADOC installation, one large U.S. Army brigade located outside of the United States, and a student sample (CAS^3) from the Army Command and General Staff College (CGSC). One of the benefits of the AZIMUTH data is that it reflects heterogeneity in source due to its administration dozens of times, to thousands of leaders, in the U.S. as well as internationally. Unfortunately, this is also one of the drawbacks of the AZIMUTH data. To preserve anonymity, information was not collected that could help describe the sample (e.g., nationality, race, age, gender, rank, etc.). While not having this information limits some conclusions, missing demographic information is somewhat common and not problematic if not making causal statements or generalizations to different populations (Craig & Hannum, 2006).

Materials

Leadership was assessed in actual units using the Leader AZIMUTH Check Version 2.0 (Form 5996), whereas the student sample used Version 2.1 (Form 6006). The AZIMUTH is

disseminated as a booklet, which has a statement of purpose and confidentiality on the cover. The first page contains an introduction to the AZIMUTH, a reminder of anonymity, instructions on form completion, a copy of the privacy act statement, and an authorization to use responses for self-development and instrumentation development. The following pages are in scantron form and contain the leadership dimension title (e.g., executing, motivating, stability, etc.) followed by four to six questions assessing each dimension with a response scale ranging from "Extremely Good Description" to "Extremely Poor Description" of that person with an additional response option of "Have not observed". This results in a 6-point scale, with an additional missing category (i.e., unobservable). Dimensions assessed by the AZIMUTH include: communicating, decision-making, motivating, developing (e.g., "Provides opportunities to learn"), learning (e.g., "Accepting of critical feedback"), building (e.g., "Actively participates in organizational/unit activities"), planning and organizing, executing, assessing, respect, selfless service ("Places the welfare of the organization before own personal gain"), integrity, emotional stability (e.g., "Maintains calm disposition under stress"), and miscellaneous items that gauge things such as a summative rating of leader effectiveness, physical fitness, and three questions to assess the instrument such as, "The questions contained in the AZIMUTH were clear and understandable". Previous studies (e.g., Laffitte & Metcalf, 2006) reported coefficient alpha in the .9s for the entire instrument. The items from version 2.0, used in the Army units are included in Appendix A. Version 2.1, used in the classroom setting among peergroups had basically the same items, however they were worded for the particular circumstance (e.g. ' your staff group' rather than 'your unit').

Procedure

There were three samples used for this exploration. The first two were from (1) an Army brigade that had two LTC battalion commanders and 3-4 GS-14 Division Chiefs and (2) a senior staff and command group (GS-14-15, O-5-7; Senior rater input from O-8). In the Army brigade data collection, the commander recommended that his direct subordinates use the AZIMUTH throughout their respective organizations and the S-1/HR group coordinated the data collection including selection of raters. In the Senior Staff and Command data collection, the IG office selected raters based on organizational charts. The third sample was from the Combined Arms Senior Staff School (CAS^3), a now defunct resident course designed for Army Captains who had completed company level command. The CAS^3 data were not true 360-degree feedback ratings because evaluations were only obtained from peers (fellow classmates), not from actual superiors and subordinates. The classroom administration also meant that participants had little chance to observe all of the behaviors rated in AZIMUTH and thus many ratings were presumably extrapolated from general impressions of the target ratee. Directions provided a consistent frame of reference, "Do not try to compare to some absolute 'ideal' standard; instead, think of others that you know of similar rank or position and use them as your 'standard' for rating."

Results

All analyses were conducted for each rating source (i.e. self, peer, subordinate, and superior) separately. A consistent pattern of results was found. All analyses were conducted aggregating across the four sources because of the consistent findings and evidence supporting conceptual equivalence. If conceptual equivalence had failed, or if a different pattern of results had been observed at the source level, then separate analyses would have been conducted and presented. Data from the TRADOC installation and the large Army Brigade located outside the United States were combined and two random samples of close to equal size were extracted. The sample was split to allow for cross-validation. This helps enhance the generalizability of the findings to the Army as a whole, rather than reflecting unit specific idiosyncrasies. Unless otherwise reported, analyses were conducted based on the first calibration sample (N = 354). Confirmatory procedures were repeated on the cross-validation sample (N = 364) as well as on the student sample (i.e., CAS^3, N = 2619).

Although the full instrument contains 72-items, not all items were analyzed. Specifically, 7 items were single item measures that were used as criteria, or to assess the entire process, and therefore were not analyzed at this stage because they did not reflect actual behavioral-based leadership dimensions. In addition, 21 negatively worded items were not analyzed because 360-degree feedback literature suggests using only positively worded items (Karrasch, et al., 1997), and subsequent versions of the AZIMUTH (after Versions 2.0 and 2.1) only contain positively worded items. In addition, negatively worded items have been found to relay different information and distort the factor structure (King, Fogg, & Downey, 2005; Schriesheim & Eisenbach, 1995). Simply put, factor analysis will clump together items that have similar distributions. Even if all items measure the same latent trait, commonly endorsed items will form distinct factors from less commonly endorsed items (Nunally & Bernstein, 1994). A separate analysis using all 65-items was conducted after the main analysis to examine the effects of negatively worded items and the factor structure variance. Overall, the same general pattern of factor structures emerged.

Descriptives

Descriptive statistics were employed to serve as a starting point for the 13 leadership dimensions, as they were originally conceptualized. Mean ratings, standard deviations, and internal consistency of all the unit ratings (N = 718) across all four rating sources are presented in Table 1. Table 2 provides detailed descriptives of the 13 leadership dimensions after removing the negatively worded items.

Table 1

Descriptive Statistics of Conceptual AZIMUTH Leadership Dimensions

Scale	Items	Alpha	Mean	SD
Assessing	5	.80	5.09	.75
Building	5	.87	5.17	.71
Communicating	6	.80	5.11	.73
Decision-making	5	.82	5.08	.81
Developing	5	.77	5.05	.76
Executing	5	.81	5.24	.63
Integrity	5	.85	5.38	.80
Learning	5	.85	5.03	.82
Motivating	5	.92	4.96	.91
Planning and Organizing	5	.79	5.14	.76
Respect	4	.84	5.33	.83
Selfless Service	5	.84	5.33	.81
Stability	5	.88	5.08	.96
AVERAGE	5	.83	5.15	.79

Table 2

Descriptive Statistics of Conceptual AZIMUTH Leadership Dimensions Without Negative Items

Scale	Items	Alpha	Mean	SD
Assessing	3	.83	5.00	.83
Building	5	.87	5.17	.71
Communicating	4	.81	5.04	.77
Decision-making	3	.83	4.96	.87
Developing	3	.78	4.94	.84
Executing	4	.85	5.10	.74
Integrity	3	.86	5.31	.83
Learning	3	.83	4.93	.85
Motivating	5	.92	4.96	.91
Planning and Organizing	3	.84	5.00	.88
Respect	3	.81	5.24	.80
Selfless Service	2	.81	5.28	.80
Stability	3	.85	4.95	1.00
AVERAGE	4	.84	5.07	.83

In short, the descriptive statistics indicated that all leadership dimensions were reliable, although there was little variability among the ratings as indicated by the relatively high value for the average (*Mean* = 5.07; theoretical maximum = 6.0), and by the relatively low amount of average variance (*SD* = .83). This indicated that although Soldiers rated consistently across items and dimensions, they did so because most people received high ratings. There was little difference between the dimensions made up of the original items and the original 13 dimensions with the negatively worded items removed. In order to test hypothesis one that there should be a two- to four-factor solution reminiscent of the original Ohio State dichotomy, a factor analysis was conducted.

Factor Analysis

Principal Components Analysis (PCA) is a technique used to identify the minimum number of factors that account for maximum variance in a data set. This technique was particularly appropriate in the present analysis because it is unaffected by multicollinearity (i.e., strong between dimension correlations). Factor extraction criteria were established a priori consistent with statistics literature (e.g., Crocker & Algina, 2006; Tabachnick & Fidell, 2006; Zwick & Velicer, 1986). Specifically, for initial extraction the factor had to have an eigenvalue greater than 1 (Kaiser, 1960). This was only the starting point, because Monte Carlo studies have indicated that this frequently used rule of thumb is "very likely to provide a grossly wrong answer", which "seems to guarantee that a large number of incorrect findings will continue to be reported" (Zwick & Velicer, 1986, p. 439). Although common, the scree plot test (Cattell, 1966) tends to overestimate the number of factors to retain and its interrater reliability among analysts is only moderate (Zwick & Velicer, 1986). The recommended procedures for ultimately deciding factor retention is Horn's (1965) parallel test (Zwick & Velicer, 1986) and Velicer's revised MAP test (Velicer, Eaton, & Fava, 2000). SPSS was used to run all PCAs, and includes the scree plot. O'Connor's (2000) add-in programs were used for Velicer et al.'s (2000) revised MAP and Horn's (1965) parallel test.

Velicer's (1976) technique is used to find the best factor solution. The MAP (minimum average partial) test ensures that it does not retain factors that have low loadings. This approach generates a PCA with only one component and calculates the average partial correlations, repeating until the solution yields the minimum average partial correlations (MAP), or values off of the main diagonal. This technique has been revised (Velicer, et al., 2000), but the basic approach remains the same. The benefit of this iterative approach is that it makes a specific decision with regard to factor extraction that is unaffected by the options selected by the analyst. Horn (1965) designed the parallel test to discover the appropriate factor cutoff. In the parallel test, a fictitious random dataset based on the same characteristics as the original (i.e., sample size, number of variables included in factor analysis) is generated and a factor analysis is conducted on the parallel, or recently generated data. Eigenvalues computed with the parallel data (in this case 1,000 permutations) are then compared to the eigenvalues generated from the original set. If a factor's corresponding eigenvalue is greater than those obtained from the parallel analysis then the factor is considered potentially important. This approach can be viewed computationally as a Monte Carlo simulation process because expected eigenvalues are obtained by simulating normal random samples that are based on the actual data set. The analyst

can be confident that the number of factors is not due to chance artifacts unique to the original data because the cut-off decision is based on literally a thousand datasets.

Consistent with Stephens (1992; as cited in Tabachnick & Fidell, 2006), reliability of the factors were then evaluated by assessing if each factor had at least four loadings > |.60|, or if each factor had at least ten loadings > |.40|. Consideration was also given to number of cross-loadings, variance accounted for, number and proportion of nonredundant residual's > |.05|, and logical fit. Factorability was assessed by examining the correlation matrix, Bartlett's test of sphericity, and Kaiser-Meyer-Olkin's (KMO) measure of sampling adequacy. The correlation matrix produced several sizable correlations and Bartlett's test of sphercity was significant, χ^2 (946) = 13112.98, $p < .0001$. Finally, Kaiser-Meyer-Olkin's (KMO's) measure of sampling adequacy was > .6 with a value of .97 indicating that the degree of common variance among the variables was more than sufficient. Before interpreting the results, a loading criterion established that only loadings >.32 (10 % shared variance) would be evaluated. In other words, weak loadings (i.e., < .32) were suppressed to ease interpretation. The initial extraction indicated a 4-factor solution accounting for 64.59 % of the variance, with 131 (13 %) nonredundant residuals with absolute values greater than .05. Although the SPSS program itself extracted four components (using only the criterion of eigenvalue > 1), the scree plot, MAP test, and parallel test all indicated a two-factor solution. It should be noted that in the original 4-factor solution, the results were uninterpretable due to a high amount of cross-loadings.

As a result of a large number of cross-loadings, previously identified multicollinearity between the items, and past ARI research efforts suggesting highly correlated subscales (e.g., Laffitte & Metcalf, 2006) the oblique rotation method was selected. In general, oblique rotations are viewed as reflecting the real world more accurately, and facilitating a more interpretable solution (Tabachnick & Fidell, 2006). The Promax rotation was selected because it is an efficient rotation method, which simplified interpretation. The Promax rotation indicated that factor 1 accounted for 21.30 % of the variance, and factor 2 accounted for 20.18 % of the variance. The pattern matrix showed a drastically reduced number of cross-loadings (down to 4) and a clear pattern similar to the Ohio State dichotomy representing task-based and interpersonal-based competencies. Unfortunately, the motivation subscale was less clear, with two items cross-loading and two items mainly loading on factor 2. This is not necessarily alarming from a theoretical perspective, because motivation, as the AZIMUTH defines it, includes creating a supportive work environment, inspiring personal excellence, and setting clear performance expectations.

Hypothesis 1 was further examined by cross-validating the PCA findings using confirmatory factor analyses. In other words, to verify consistent findings across samples, a different technique assessed the degree to which the original findings were replicated across samples. A higher-order confirmatory factor analysis with assessing, building, communicating, decision-making, developing, executing, learning, planning and organization driving the *task competency factor*, and, integrity, respect, selfless service, and stability driving the *interpersonal competency factor* was run for all three samples (see Table 3). This analysis was repeated with all leader dimensions forced onto a single competency, for comparison purposes.

Table 3

Comparison of Confirmatory Factor Analytic Models

Model	CFI	RMSR	RMSEA	N
S1 Two-Factor Model	.884	.048	.073	354
S1 Single-Factor Model	.868	.053	.077	354
S2 Two-Factor Model	.896	.048	.072	364
S2 Single-Factor Model	.886	.049	.075	364
CAS^3Two-Factor Model	.931	.022	.050	2619
CAS^3 Single-Factor Model	.923	.023	.053	2619

S1 = half original sample, S2 = cross-validation from other half of original sample, CAS^3 = cross-validation from student sample

Table 3 shows that the all models at least marginally fit the data. Both single- and two-factor models were best fitted to the student sample (i.e., CAS^3). In all cases, the two-factor solution provided the best fit; however, the gains over the single-factor model were minimal. Thus, hypothesis one, predicting 2-4 factors, received somewhat mixed support. In order to review hypothesis two, in which the number of raters required to produce consistent ratings were predicted to vary by source, intraclass correlations (i.e., reliability analysis) were analyzed.

Reliability

Intraclass correlations (1, 1) and (1, k) based on a one-way Analysis of Variance (ANOVA) in which the leader being rated is treated as a random effect and rater is treated as measurement error were conducted for all original leadership dimensions. In other words, the agreement among the raters from each source were analyzed while taking into account that some raters may rate several target leaders, and others might rate only one target leader. Inter-rater reliability estimates for peer, subordinate, and superior ratings of the two factors are provided in Table 4. Inter-rater reliability estimates of the different rating sources for the original thirteen dimensions are presented in Appendix B. The number of raters required to produce convergent ratings (i.e., > .69) are presented in Appendix C.

Overall, it usually took just a single rater to provide reliable (i.e., >.69 convergence) ratings. In practice, the AZIMUTH could be reliably and confidentially assessed with three raters. All administrations of AZIMUTH to date have used three raters as the minimum number of subordinate and peer raters because that is the threshold at which reasonable anonymity can be preserved. Consistent with the second hypothesis, there was some evidence of differential reliability, with superiors providing the most consistent ratings; however, the difference was so small that it was not practically meaningful.

Table 4

Intraclass Correlations for Ratings of Each Rating Source

	Peer	Subordinate	Superior
Task Competency			
(ICC 1, 1)	.71	.71	.74
(ICC 1, k)	.95	.95	.96
Interpersonal Competency			
(ICC 1, 1)	.64	.63	.72
(ICC 1, k)	.88	.87	.91

ICC (1, 1) denotes the form where for each subject, one randomly samples from the rater pool k different raters to rate this subject. Therefore, the raters who rate one subject are not necessarily the same as those who rate another. This design corresponds to a one-way Analysis of Variance (ANOVA) in which subject is a random effect, and rater is viewed as measurement error. The reliability is calculated from a single measurement. ICC (1, k) denotes the same as above, but reliability is calculated by taking an average of the k raters' measurements.

Additional Analyses

Additional analyses were conducted in response to the factor analytic and construct validity results in an effort to identify the utility of the 2-competency factor structure in the prediction of single item measures of "maintains effective interpersonal relations with others", "is effective on the job", and "this person is a good leader". Based on previous leadership literature and the AZIMUTH development process (Blake & Mouton, 1964; Fleishman, 1953; Mumford et al., 2007; Laffite & Metcalf, 2006), it was hypothesized that:

Hypothesis 1a. The two competencies would b*e strongly positively related, but not reach singularity*

Hypothesis 1b. The two competencies would c*onverge with single-item assessments of maintaining interpersonal relations, job effectiveness, and overall assessment of leadership.*

Hypothesis 1c. The interpersonal competency should better predict the maintaining interpersonal relations criterion.

Correlations between the two competencies were quite high (sample 1 r = .83, sample 2 r = .84, sample 3 r = .79). After correcting for disattenuation because of unreliability, the two competencies approached singularity (sample 1 r = .90, sample 2 r =.90, sample 3 r =.91). In other words, the two competencies had over 80 % shared variance and appeared to measure virtually the same thing. Thus, Hypothesis 1a received mixed support in that the two competencies were strongly positively related, but they also were close to reaching singularity. In other words, there was support for convergent validity, but not discriminate validity. Both convergence and discrimination are common forms of evidence regarding construct validity. Regression analysis indicated that both competencies combined accounted for the majority of

variance in all three criteria. Specifically, in the first sample 59 % of the variance in "maintains effective interpersonal relations with others" was accounted for by both competencies. Consistent with expectations the interpersonal competency did have a higher beta (β = .42) than the task competency (β = .39), when both competencies were simultaneously entered. The zero-order relationships; however, indicated almost the exact same strength between the task competency and maintaining effective interpersonal relations (β = .73) and the interpersonal competency and maintaining effective interpersonal relations (β = .74). Both competencies also accounted for the majority of variance in job effectiveness, although the *task competency*, was relatively more important (β = .63) than the *interpersonal competency* (β = .14). The same pattern emerged for predicting the single-item measure that "this person is a good leader". Again, both competencies combined to account for the majority of the variance (71 %), with task competency being relatively more important (β = .62) than the interpersonal competency (β = .26). Similar results were obtained in both the secondary sample and the student sample. The results of all regression analyses are presented in Table 5.

Table 5

Regression Coefficients Predicting Overall Leadership

Sample	Criterion	% variance R^2	Zero-order C1, C2	β in model C1, C2
S1	Maintains effective interpersonal relations	59 %	.73, .74	.39, .42
	Is effective on the job	57 %	.75, .66	.63, .14
	Is a good leader	71 %	.83, .77	.62, .26
S2	Maintains effective interpersonal relations	63 %	.77, .76	.44, .39
	Is effective on the job	57 %	.76, .66	.70, .07
	Is a good leader	74 %	.85, .79	.63, .26
CAS^3	Maintains effective interpersonal relations	52 %	.71, .65	.50, .26
	Is effective on the job	57 %	.75, .64	.66, .11
	Is a good leader	58 %	.75, .67	.58, .21

C1 = *task competency* (assessing, building, communicating, decision-making, developing, executing, learning, planning and organizing), C2 = *interpersonal competency* (integrity, respect, selfless service, stability). S1 = half-original sample, S2 = cross-validation from other half of original sample, CAS^3 = cross-validation from student sample.

The third hypothesis questioned the conceptual similarity of the ratings (i.e., were the same dimensions considered equally important across self, peer, subordinate, and superior rating sources). The large amount of variance accounted for in the three criteria using the two competencies supports Hypothesis 1b, thus further providing support for the construct validity of the AZIMUTH. At the same time, the expected, but unobserved differential prediction of maintaining effective interpersonal relations is further indication of a lack of discrimination between the two competencies.

Conceptual Agreement

Conceptual agreement was tested in accordance with Cheung's (1999) multifaceted conception using the same multitrait-multirater (MTMR) confirmatory factor analytic (CFA) approach used by Woeher et al. (2005). First, a comparison model was evaluated, which ignored rating source effects and only looked at leadership dimension loadings. Second, a factorial or configural invariance model was tested. Third, a metric invariance model was tested. In the factorial invariance model, the relative importance of items as indicators is tested (i.e., are items more predictive of a particular competency because of the rating source?). To test factorial equivalence the model specified that all ratings (regardless of dimension) loaded on their respective performance dimension and specified that the ratings loadings from all four sources to be unconstrained and the unique variances of the ratings are unconstrained. A lack of model fit would indicate that ratings of each leadership dimension measure different dimensions depending on the rating source. If this first MTMR-based factorial equivalence model fits the data and showed a marked improvement over the dimension-factor-only comparison model then a metric invariance model is tested. In the metric invariance model, factor loadings for each leadership dimension are set equal across rating sources (i.e., constrained). A lack of model fit would indicate that item loadings vary (i.e., are not equivalent) across rating sources. Confirmation of both models is a sufficient condition for conceptual equivalence (Cheung, 1999; Woeher et al., 2005). Results of model testing for conceptual agreement are presented in Table 6.

Table 6

Results of Conceptual Equivalence Model Testing

Model	χ^2	***df***	$\Delta\chi^2$	Δdf	***RMSEA***	***CFI***
Leadership dimension factors only	375.20	20			0.21	0.57
Configural invariance dimensions and rating source factors	359.4	13	15.80	7	0.02	1.00
Metric invariance dimensions and rating source factors	339.22	12	20.18	1	0.04	0.99

Results indicated that the comparison model poorly fit the data, while the configural invariance model had the best fit, and the metric invariance model had acceptable fit. The acceptable levels of fit for both MTMR models (i.e., configural and metric invariance), as well as the significant improvement over the leadership dimension-factors-only model supports the conceptual agreement of the AZIMUTH across rating sources. This supports the third hypothesis. In other words, each rating source had the same mental model, which allowed for accurate between source comparisons ('apples compared to apples'). The final analysis examined rating patterns.

Rating Patterns

The estimator approach (Van Velsor, Taylor, & Leslie, 1993; Yammarino & Atwater, 1993) is a useful technique for evaluating self-awareness. The estimator approach creates a difference score by taking self-ratings for a given construct or scale, and subtracting them from another rating, from a different source (i.e., peers, subordinates, superiors) on the same construct or scale. Difference scores are then evaluated to the newly created domain of difference scores and are classified depending on the individual difference score's relation to the distribution of the domain of difference scores, usually by a pre-set standard deviation criterion such as within one standard deviation (Roush & Atwater, 1992; Van Velsor, et al., 1992). This method assesses individual estimation patterns of one's own performance in a comparatively inflated or modest manner. The feedback provided to the individual leader can describe their pattern, allowing the individual to see some 'evidence' of how others see them compared to how they view themselves.

Table 7 reviews distributional properties for the four rating sources. All ratings were collected using a 6-point rating scale, and the lowest mean rating was 5.05. Self-ratings produced the lowest mean, least amount of negative skew, and least variance. Thus, Hypothesis 4, which stated that self-ratings would be higher than the ratings from other sources, was not only disconfirmed, but the exact opposite appeared to be the case. Superiors had the most amount of skewness, and most variance. The estimator approach described earlier was used to classify self-other perceptions as accurate, under-, or over- estimators on the basis of absolute agreement (i.e., an identical response).

Table 7

Distributions of Self, Peer, Subordinate, and Superior Ratings

Rater	Measure	*M*	*SD*	Skewness
Self	Task Competency	5.05	.45	-.03
	Interpersonal Competency	5.34	.42	-.20
Peer	Task Competency	5.15	.58	-.11
	Interpersonal Competency	5.28	.67	-.37
Subordinate	Task Competency	5.16	.58	-.13
	Interpersonal Competency	5.34	.64	-.67
Superior	Task Competency	5.08	.76	-.18
	Interpersonal Competency	5.17	.89	-.45

Standard error of skewness = .287. Based on 6-point Likert scale.

Tables 8 and 9 show that for any given comparison of task competency the vast amount of officers provided ratings in agreement with others' ratings. The trend observed was that there was a similar proportion of over-estimators and under-estimations. Regardless of the measure, self ratings were more in agreement with peers than with superiors. Self-superior differences were the largest; however, identical responses still occurred over 80 % of the time for both task and interpersonal competencies. Virtually identical results were obtained for the estimation accuracy of task competency and interpersonal competency. In sum, there was a large degree of self-other agreement.

Table 8

Frequency Analysis of Estimation Patterns of Task Competency

Self Compared With	Agreement	Under	Over
Peers	(N= 371) 87.30%	(N = 29) 6.80%	(N = 25) 5.90%
Subordinates	(N = 380) 89.40%	(N = 26) 6.10%	(N = 19) 4.50%
Superiors	(N = 342) 80.50%	(N = 39) 9.20%	(N = 44) 10.40%

Table 9

Frequency Analysis of Estimation Patterns of Interpersonal Competency

Self Compared With	Agreement	Under	Over
Peers	(N= 373) 87.80%	(N = 28) 6.60%	(N = 24) 5.60%
Subordinates	(N = 381) 89.60%	(N = 24) 5.60%	(N = 20) 4.80%
Superiors	(N = 344) 80.90%	(N = 31) 7.30%	(N = 50) 11.80%

Discussion

The purpose of 360-degree feedback is to provide valuable information from different sources. The present work investigated AZIMUTH rating patterns. Before feedback is shared with the ratee (i.e., target leader) the appraisal should be assessed for differences between raters within a source (i.e., establishment of inter-rater reliability), and conceptual equivalence between sources (Greguras & Robie, 1998). Distributions of ratings, factor structures, reliability, conceptual equivalence between sources, and estimation patterns produced using the AZIMUTH were all examined. Initial descriptive statistics showed that both leadership common competencies were rated consistently (i.e., were reliable), and there was a reduced amount of variability among the ratings due to most target leaders receiving a rating around a five on the six-point scale. The results were positive and support the AZIMUTH as a reliable instrument.

Furthermore, the present research extends the recent findings of Laffite and Metcalf (2006), and indicates that the AZIMUTH is in fact well on its way to the intended goal of being a generic "one-size-fits-all" leader assessment.

Given the pervasiveness of 360-degree feedback systems (Edwards & Ewen, 1998), and the continued need to create a more comprehensive typology of leadership skills (Connelly et al., 2000; Kanungo & Misra, 1992; Wright, 1996; Wright & Taylor, 1985; Yukl & Van Fleet, 1992), and the importance of enhancing military leadership (Marvin, 1995; Wong, et al., 2003) this research focused on creating a theoretically rich conceptual model of leadership common competencies. Validating the common competency model is a necessary and important step to understanding assessment and development of the higher-order strategic leadership competencies described by Wong, Gerras, et al. (2003).

Validity evidence was investigated consistent with guidelines provided by the American Educational Research Association, American Psychological Association, National Council on Measurement in Education (1999), and the Society for Industrial/Organizational Psychology (2003). Four relevant sources of evidence were investigated namely, content-related evidence, internal structure evidence, construct-related evidence, and conceptual agreement across rating sources. Content validation was supported by the developmental process, which defined the Army leadership construct using scientific literature and Army leadership and leader development information. This process also utilized subject matter experts from a variety of sources (both internally and externally), along with rigorous pilot testing and reevaluation by different subject matter experts.

Internal structure evidence was garnered by both an exploratory and confirmatory factor analysis to test Hypothesis 1, that a two to four factor solution incorporating the original Ohio State dichotomy would fit the data. The initial model did indicate a 2-factor solution similar to the initiating structure and individual consideration factors from The Ohio State Leadership Studies and The Leadership Grid (e.g., Blake & Mouton, 1964; Fleishman, 1953), thus offering support of the first hypothesis. Additionally, this 2-factor solution fit the content validation structure in which task competency items and interpersonal items loaded on their respective leadership competency.

Specifically, the task-based competency was compromised of related constructs of assessing, building, communicating, decision-making, developing, executing, learning, and planning and organizing. The Interpersonal-based competency was comprised of related constructs of integrity, respect, and selfless service. Confirmation of this factor structure occurred in both a random sample comprised of the other half of the original factor, and an entirely different sample of student ratings. While hypothesis 1 initially appeared to be strongly supported, comparisons with a single-factor solution showed a modest gain using the dichotomy. It is likely that the two-factor solution is the best fit, and its value appears minimized due to a strong general factor. This general factor may be due to either true or artificial halo, in that ratings from a source across dimensions were virtually the same regardless of the rating source or target leader. In sum, the AZIMUTH data did not indicate much discrimination in terms of leadership dimensions.

Inter-rater Reliability. Hypothesis 2 stated that the number of raters required to produce consistent ratings would vary by source. Inconsistent with previous findings (e.g., Conway & Huffcutt, 1997; Greguras & Robie, 1998), and in contradiction to Hypothesis 2 there was little disparity in rating consistency (i.e., differential rater reliability) depending on the rating source and the leadership competency being evaluated. Although superiors tended to produce the most consistent ratings, their consistency was virtually the same as the other sources. There was a greater difference in what was being rated (i.e., task or interpersonal competency) than who was rating. Examining the bigger picture, it appears that when using the AZIMUTH superiors, subordinates, and peers do in fact give reliable ratings.

Additionally, there were a relatively smaller number of raters required to produce reliable ratings. This means that a reliable 360-degree Army feedback system may need only about 3 peers, 3 subordinates, and 3 superiors, plus self rating. This estimate of 10 necessary raters is much more favorable than other estimates, which are as high as 23 raters including self (Greguras & Robie, 1998). However, each competency in the present analysis was assessed with roughly seven times the number of items that Greguras and Robie used to develop their estimates. These findings also suggest that despite the personal biases found in Army personnel (Marvin, 1995; Karrasch & Halpin, 1999), peers do in fact provide reliable ratings, and that no source is so much more reliable that their responses can be used as the gold standard. Having established reliability of the AZIMUTH, the next step was to examine the construct validity evidence.

Construct Validity. Construct validity support was gained from a series of correlation and regression analyses in which the two competencies were expected to be strongly positively related, but not reach singularity (Hypothesis 1a), and converge with single-item assessments of maintaining interpersonal relations, job effectiveness, and overall assessment of leadership (Hypothesis 1b). Finally, the interpersonal value-based skills subscale was expected to correlate more strongly with the maintaining interpersonal relations criterion (Hypothesis 1c). The two subscales were strongly correlated ($r = .83$); however, they began to approach singularity after correcting for disattenuation because of unreliability ($r = .90$). This observed correlation was somewhat higher than anticipated given recent evidence (Mumford et al., 2007).

Hypothesis 1b was strongly supported in that the range of correlations of the two factors and single-item measures of maintaining interpersonal relations, job effectiveness, and overall assessment of leadership were within in the .6s (*Range* = .64 to .85). In all cases, combining both competencies accounted for the majority of the variance in the criteria. While the competencies as a set predicted the criteria well, the lack of differential prediction between the factors and "Maintains effective interpersonal relations with others" did not support Hypothesis 1c. In addition, it appeared that regardless of the criterion usually the best predictor was the *task competency*. This does not mean that the *interpersonal competency* has no value, because in all cases the prediction of the criteria was enhanced by the addition of the second factor. In other words, using either competency will yield a good prediction of the leadership criteria used in this analysis, and using both competencies will yield even greater predictive power, but when combined the *task competency* was relatively more important than the *interpersonal competency*. Thus, there is some evidence of construct validity in that the competencies were related to leadership effectiveness, but the evidence also suggests the factors do not discriminate, as was

expected. At the crux of inferring inter-rater reliability and validity information for the interpretation of AZIMUTH results is the issue of conceptual agreement.

Measurement Equivalence. Conceptual agreement is an assumption often made, but rarely tested (Bagous, 2004; Byrne, Shavelson & Muthen, 1989); however, Lance and Bennett (1997) show that this is a dangerous assumption. Lance and Bennett (1997) evaluated self, superior, and peer ratings of eight samples of U. S. Air Force airmen and found a lack of equivalence in five out of the eight Air Force airmen samples. A major advantage of the MTMR CFA method is testing between rating source differences simultaneously. While previous measurement equivalence for classroom data based on self and classmate Captain assessments had been recently established (Laffitte & Metcalf, 2006), the test of whether or not the AZIMUTH is a generic "one-size-fits-all" instrument had not been previously tested. The present research supports Hypothesis 3 that Army officers provide conceptually equivalent ratings. In other words, the dimensions of leadership and relative importance were considered conceptually the same regardless of the source who is evaluating. Knowing that the instrument produced reliable ratings, with a consistent factor structure in line with prevailing leadership research, and conceptually equivalent across rating sources, inspires confidence that inferences drawn from the AZIMUTH are valid. Therefore, the final exploration examined rating patterns, in terms of between rating source agreement and self-severity bias.

Rating Patterns. Agreement between rating sources has been described several different ways in the literature. For example, Vance and colleagues (1988) stated that convergence was an indicator for construct validity, whereas Atwater and Yammarino (1992) described convergence as self-awareness, and Yammarino & Atwater (1993) later referred to convergence as accuracy. Mersmen and Donaldson (2000) clarified that convergence alone could merely indicate shared biases. Regardless of the value convergence is given, or even the level of desirability, rating patterns in general are important and raise interesting research questions (Mersmen & Donaldson, 2000). It should be noted that others (e.g., Edwards, 1994) suggest polynomial regression analyses, instead of difference scores. While polynomial regression is generally a preferred method, it is most appropriately applied to situations of conceptual inequivalence and when there is an external objective criterion (not just another rating from a different source). The present dataset does not contain an external objective criterion, and the present research demonstrated measurement or conceptual equivalence between the rating sources.

Self-ratings produced the lowest mean, least amount of negative skew, and the least variance. Thus, Hypothesis 4, which stated that self-ratings would be higher than the ratings from other sources, was not only disconfirmed, but the exact opposite appeared to be the case. The present findings are inconsistent of general North American organizational studies, but consistent with the military literature, in that Army officers indicated a modesty bias regardless of what was being rated. Past research has demonstrated a strong tendency for self-ratings to be higher than other rating sources (Fox & Dinur, 1988; Harris & Schaubroeck, 1988; Landy & Farr, 1980, London & Wholers, 1989; Mabe & West, 1982; Podsakoff & Organ, 1986; Thornton, 1980). However, a modesty bias has also been found in other military contexts, with Naval officers (Bass & Yammarino, 1991) and with Air Force officers (e.g., Halverson, et al., 2002). Therefore, findings of the present research were congruent with previous military findings (Bass

& Yammarino; Conrad, 2004; Halverson et al.), but contrary to general organizational literature (e.g., Harris & Schaubroeck, 1988; Podsakoff & Organ, 1986).

This suggests that in the Army, and at least some other branches of the military, it is either less acceptable to self-rate highly, or it is more acceptable to rate others highly. What is especially interesting in the context of the present research is that while the average self-ratings were lower than the average ratings from the other sources, the self-other difference scores had a mean near 0, and nearly all of the self-other ratings were within one standard deviation. In fact, over 80 % of the leaders surveyed were in total agreement in their self vs. other ratings. In other words, self-raters indicated a modesty bias (i.e., lack of leniency directed at self) and other sources indicated a leniency bias. In addition, self-peer agreement was higher than self-superior agreement. Again, this suggests that peers provide reasonably consistent information.

Differences between military branches (e.g., Army vs. Air Force) are present and can be expected. For example, Halverson et al. (2002) indicated that although self-ratings of leadership were lower than superiors and subordinates, they were higher than peer ratings. Such differences can be expected because each military branch has, "a unique culture that influences acceptable leadership styles in that service" (English, 2002, p. 3). Thus, one explanation why the Army has a modesty bias is that there are cultural norms in day-to-day behavior, in evaluation of others, and in the conceptualization of leadership that influence the 360-degree feedback process (Blanton & Christie, 2003).

Farh, Dobbins, and Cheng (1991) previously demonstrated that Taiwanese workers also had a modesty bias. A comparison of nations by Hofstede (1980) indicated that Taiwan was among the highest countries in Confucian Work Dyanism, which is a collection of work attitudes that stress order, steadiness, and respect for the social hierarchy. These values, which are especially espoused by the Taiwanese are also espoused in Army. This may indicate why both groups have self-ratings that reflect a modesty bias. While Farh, et al.'s (1991) explanation that the observed modesty bias was simply the product of broad cultural factors (i.e., East vs. West) appeared inadequate by subsequent research (Jiayuanyu & Murphy, 1993), it does appear that more specific subcultural factors (i.e., Taiwanese workers vs. other eastern workers or Army vs. Air Force) could help explain this phenomenon.

Unfortunately, this hypothesis was not tested and many other plausible explanations remain. For example, another possible explanation is that these estimation patterns are manifestations of personality traits. Research has shown that those who are low in self-esteem (Farh & Dobbins, 1989) or high in self-consciousness (Nasby, 1989) tend to exhibit under-rater estimation patterns like those observed in the AZIMUTH data. An alternative argument provided by Mersman and Donaldson (2000) is that underrating or modesty may be a function of how the other raters were selected. Specifically, they argued that when the target leader (ratee) selects and has a high need for approval, there is a tendency for modesty bias. Unfortunately, specific data collection procedures were not coded, and do not allow for a direct test of this hypothesis.

Limitations

Before concluding, it is important to address the limitations of this work. This work faces the same challenges as other archival investigations. Specifically, data were already collected and therefore some information that would help contextualize the findings were missing such as a description of the source of data (unit, rater rank or position, ratee rank or position, duration and quality of rater-ratee relationship). Additionally, individual characteristics for both rater and ratee (age, sex, and training) and the selection of raters (selecting raters who were convenient versus raters who had opportunities to observe leadership behaviors, versus raters who were in a position to be more objective) were not available. We do know the criteria by which raters were to be chosen in each of the two unit/organizational samples (i.e., "persons who work with the target leader are who should be in a position to rate their leadership") but we don't know the extent to which the raters in fact were familiar with the ratees. In the student sample, class instructors assigned ratees to peer raters using a random process that equalized the workload but little if any attention was given to pairing those who had, for example, frequently worked together on class projects. Finally, although highly desired, external criteria such as promotion records, or objective test scores were unavailable.

It is most important that criteria be relevant, reliable, and uncontaminated (SIOP, 2003). The criteria in this investigation were certainly relevant in that they reflected various groups' (superiors, subordinates, peers, and leader) perceptions of aspects important to Army leadership. The single-item nature of these criteria gives pause, because single-item measures do not allow for specific reliability testing, and by definition afford larger measurement errors than composite criteria (Nunnally & Bernstein, 1994). However, each single-item measure was rated by three to twelve raters for each leader, there was a large sample size, and if the three criteria were combined to create a single criterion the correlation pattern would remain the same, and the composite would be internally consistent ($\alpha = .87$). Contamination concerning location, the leader's unit, rater characteristics, and rater-ratee relationship was expected to be minimal because the sample was derived from a variety of leadership settings, using several large Army units, in which the rater selection varied. A bigger source of criteria contamination was in the raters' knowledge of subscale scores. That is, the single-item measures were assessed after the raters had already rated the task and relationship skills. Therefore, it would be expected that the relationships between the two competencies and the criteria were probably inflated. While this raises some caution in interpreting the construct validity evidence presented, the analysis of rating patterns indicated that self-other sources were in absolute agreement over 80 % of the time.

Despite these apparent limitations, specific causal models were not being tested, so this analysis does not suffer from a third variable, or lack of control problem. Additionally, age, rank, unit, and rater selection, among other variables, while important, can be thought of as somewhat controlled since the analysis combined two units, randomly sampled each half, and confirmed findings with an entirely different military sample (i.e., CAS^3 students), and therefore the present findings are unlikely to be mere artifacts of instructional biases, or unit idiosyncrasies.

Conclusions

Because of the present work in building a valid model of common competencies relevant to Army officers, the next long-term goal of this project would be to refine these competencies at the distinctive level (Brownell, 2006) taking into account contingencies (Blake & Mouton, 1964). In other words, after further support, this model could become more sophisticated and useful by incorporating adaptability elements. The present research project has several implications for both researchers and practitioners. First, this analysis answered some of the essential unanswered questions regarding the psychometric qualities of the AZIMUTH. Second, this analysis established that a single rater from each source fulfilled acceptable reliability; however, it would likely take an additional two raters from each source to ensure acceptable anonymity. This finding is also important because it suggests that 360-degree feedback ratings in the Army may be more reliable than previously thought. The suggestion of a smaller rater number requirement obviously reduces the entire strain necessary to implement a 360-degreefeedback program.

Third, the finding that peer inter-rater reliability was as generally as good as more preferred sources (i.e., subordinates and superiors) and the finding that self-peer ratings were more often in agreement than self-superior ratings is interesting. Although Army officers reacted least favorably to being evaluated by peers in terms of informational value, appropriateness, and accuracy (e.g., Karrasch & Halpin 1999), the present analysis shows that if nothing else these evaluations are reliable, conceptually equivalent to ratings from the other sources, and in absolute agreement with self-ratings over 80 % of the time. One reason that Army officers are dismissive of peer reports is that their peers are seldom in a position to view them in action as leaders (Marvin, 1995). As a result, there may be limited information available that a peer has in order to make their decision. Superior ratings also are not viewed as particularly useful (Karrasch & Halpin, 1999) because they tend to supplement and are often redundant with other on-the-job feedback. In contrast, subordinate ratings are viewed as the most useful because they are thought to provide less redundant information (Karrasch & Halpin; Halverson et al., 2002), which is interesting given that the present findings indicated that subordinate-self agreement was highest.

Fourth, the modesty bias observed in the present analysis is the opposite of Western organizational literature. Further research should explore this finding to determine if it is an artifact of this dataset. The bias could be indicative of an element of Army culture. If such a bias is stable it could provide fertile ground for extending Atwater et al.'s (1995) and Halverson et al.'s (2002) finding that accurate self-assessments are predictive of high performance, and under- and over- estimators require feedback presented to them in different ways, and have specific outcomes.

In Yammarino's (2000) evaluation of the leadership skill literature, he argued that there were four key related research areas:

- Development of reliable and valid measures
- Enhanced prediction of leader performance beyond other constructs associated with effective leadership
- Development of higher-level skills that are linked to experiential growth
- Identification of other constructs associated with patterns of leader growth

The present set of analyses has advanced the first core research area. The next research phase is to demonstrate incremental validity of the AZIMUTH skills beyond other leadership constructs. While Flowers (2004) recommended the identification of leadership competencies is also the first step in a process of creating an Army of strategic leaders, he created an applied plan. Following Flowers' recommendation, these competencies should be institutionalized, something that is already on-going. According to Flowers, a process that is based on developing confidence and enduring competencies "will lead to an Army able to win in any environment" (Flowers, 2004, p.45).

References

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, D. C.

Atwater, L. E., Roush, P., & Fischthal, A. (1995). The influence of upward feedback on self and follower ratings of leadership. *Personnel Psychology, 48,* 35-60.

Atwater, L. E., & Yammarino, F. J. (1992). Does self-other agreement on leadership perceptions moderate the validity of leadership and performance on predictions? *Personnel Psychology, 45,* 141-164.

Bagous, A. M. (2004). *Multi-source feedback: A confirmatory factor analysis approach to measurement equivalence*. Paper presented at the annual conference of the Society for Industrial and Organizational Psychology, Chicago, Illinois.

Bass, B. M., & Yammarino, F. J. (1991). Congruence of self and other's leadership ratings of naval officers for understanding successful performance. *Applied Psychology: An International Review, 40,* 437-454.

Blake, R., & Mouton, J. (1964). *The Managerial Grid: The Key to Leadership Excellence*. Houston: Gulf Publishing Co.

Blanton, H., & Christie, C. (2003). Deviance: A theory of action and identity. *Review of General Psychology, 7,* 115-149.

Bowers, D. G., & Seashore, S. E. (1966). Predicting organizational effectiveness with a four-factor theory of leadership. *Administrative Science Quarterly, 11,* 238-263.

Bozeman, D. P. (1997). Interrater agreement in multi-source performance appraisal: A commentary. *Journal of Organizational Behavior, 18,* 313-316.

Brownell, J. (2006). Meeting the competency needs of global leaders: A partnership approach. *Human Resource Management, 45,* 309-336.

Byrne, B., Shavelson, R. & Muthen, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin,* 105, 456-466.

Bunker, K. A., Kram, K. E., & Ting, S. (2002). The young and the clueless. *Harvard Business Review, 0078012, 80*(12).

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research,* 1(2), 245-276.

Cheung, G. W. (1999). Multifaceted conceptions of self-other ratings disagreement. *Personnel Psychology, 52*, 1-36.

Connelly, M. S., Gilbert, J. A., Zaccaro, S. J., Threlfall, K. V., Marks, M. A., & Mumford, M. D. (2000). Predicting organizational leadership: The impact of problem solving skills, social judgment skills, and knowledge. *The Leadership Quarterly, 11*, 65-86.

Conrad, T. M. (2004). *An item-level analysis of the Leader* AZIMUTH *Check.* Unpublished manuscript. Ft. Leavenworth, Kansas: Army Research Institute.

Conway, J. M., & Huffcutt, A. I. (1997). Psychometric properties of multisource performance ratings: A meta-analysis of subordinate, supervisor, peer, and self-ratings. *Human Performance, 10*, 331-360.

Craig, S. B., & Hannum, K. (2006). Research update: 360-degree performance assessment. *Consulting Psychology Journal: Practice and Research, 58*, 117-122.

Crocker, L., & Algina, J. (2006).*Introduction to classical & modern test theory.* Thomson Wadsworth: Mason, OH.

Edwards, J. R. (1994). Regression analysis as an alternative to difference scores. *Journal of management 20,* 683-689.

Edwards, M., & Ewen, A. (1998). *Multisource assessment survey of industry practice.* Paper presented at the 360-Degree Feedback Global Users Conference, Orlando, FL.

English, A. (2002). *The masks of command: Leadership differences in the Canadian Army, Navy, and Air Force.* Paper presented at the Leadership in the Armies of Tomorrow and the Future, Ontario, Canada.

Farh, J. L., Dobbins, G. H. (1989). Effects of self-esteem on leniency bias in self-reports of performance: A structural equation model analysis. *Personnel Psychology, 42*, 835-850.

Farh, J. L., Dobbins, G. H., & Cheng, B. S. (1991). Cultural relativity in action: A comparison of self-ratings made by Chinese and US workers. *Personnel Psychology, 44*, 129-147.

Fleishman, E. A. (1953). The description of supervisory behavior. *Personnel Psychology, 37*, 1-6.

Flowers, M. (2004). Improving strategic leadership. *Military Review, March-April*, 40-46.

Fox, S., & Dinur, Y. (1988). Validity of self-assessment: A field evaluation. *Personnel Psychology, 41*, 581-592.

Garavan, T., Morley, M., & Flynn, M. (1997). 360-degree feedback: Its role in employee development. *Journal of Management Development, 16*, 134-147.

Greguras, G. J., & Robie, C. (1998). A new look at within-source interrater reliability of 360-degree feedback ratings. *Journal of Applied Psychology, 83*, 960-968.

Guion, R. M. (1997). *Assessment, measurement and prediction for personnel decisions*. Mahwah, NJ: Lawrence Erlbaum Associates.

Halpin, S. M. (1997). *The Leader* AZIMUTH *Check: A leader self-development tool.* Unpublished manuscript, Ft. Leavenworth, Kansas: Army Research Institute.

Halverson, S. K., Tonidandel, S., Barlow, C., & Dipboye, R. L. (April, 2002). *Self-other agreement on a 360-degree leadership evaluation.* Paper presented at the annual meeting of the Society for Industrial and Organizational Psychology, Toronto, Canada.

Harris, M. M., & Schaubroeck, J. (1988). A meta-analysis of self-supervisor, self-peer, and peer-supervisor ratings. *Personnel Psychology, 41*, 43-62.

Hatfield, B. Jr. (1997). *Strategic leadership development: An operation domain application*. Research Paper (0607). Maxwell-Gunter Air Force Base, Montgomery, Alabama: Air Command and Staff College.

Hazucha, J. F., Hezlett, S. A., & Schneider, R. J. (2003). The impact of 360-degree feedback on management skills development. *Human Resource Management, 32*, 325-352.

Hofstede, G. (1980). *Culture's consequences*. Newbury Park: Sage.

Horey, J., Fallesen, J. J., Morath, R., Cronin, B., Cassella, R., Franks Jr., W., & Smith, J. (2004). *Competency based future leadership requirements* (ARI Technical Report 1148). Ft. Leavenworth, Kansas: Army Research Institute.

Horey, J., Harvey, J. Curtin, P., Keller-Glaze, H., Morath, R., & Fallesen, J. (2006). *A criterion-related validation study of the Army Core Leader Competency Model* (ARI Technical Report 1199). Ft. Leavenworth, Kansas: Army Research Institute.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30*, 179-185.

Jiayuanyu, & Murphy, K. R. (1993). Modesty bias in self-ratings of performance: A test of the cultural relativity hypothesis. *Personnel Psychology, 46*, 357-263.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement, 20*, 131-151.

Kanungo, R. N., & Misra, S. (1992). Managerial resourcefulness: A reconceptualization of management skills. *Human Relations, 45*, 1311-1332.

Karrasch, A. I. (2006). *The Army leader assessment and feedback pilot program* (CAL Technical Report 2006-1.). Ft. Leavenworth, Kansas: Center for Army Leadership.

Karrasch, A. I., & Halpin, S. M. (1999). *Feedback on 360-degree Leader* AZIMUTH *Check assessment conducted at Ft. Clayton, Panama* (ARI Research Note 99-21.). Ft. Leavenworth, Kansas: Army Research Institute.

Karrasch, A. I., Halpin, S. M., & Keene, S. D. (1997). *Multirater assessment process-a literature review* (ARI Technical Report 1076). Ft. Leavenworth, Kansas: Army Research Institute.

King, C.V., Fogg, R.J., & Downey, R.G. (2004, April). *The positives and negatives of negatively worded items in scales*. Paper presented at the annual meeting for the Society of Industrial and Organizational Psychology, Chicago, Illinois.

Laffitte, L. J., Halpin, S. M., & Tran, T. (n.d.). *The evolution of the Leader* AZIMUTH *Check*. Unpublished manuscript, Ft. Leavenworth, KS: Army Research Institute.

Laffitte, L. J., & Metcalf, K. A. (2006, May). *Optimizing factor structures with measurement equivalence using confirmatory factor analysis and item response theory*. Symposium presented at the annual meeting of the Society for Industrial and Organizational Psychology, Dallas, TX.

Landy, F. J., & Farr, J. L. (1980). Performance rating. *Psychological Bulletin, 87*, 72-107.

London, M. & Beatty, R.W. (1993). 360-degree feedback as a competitive advantage. *Human Resource Management, 32*, 353-372.

London, M., & Smither, J. W. (1995). Can multi-source feedback change perceptions of goal accomplishment, self-evaluations, and performance related outcomes? Theory-based applications and directions for research. *Personnel Psychology, 48*,803– 839.

London, M., & Wohlers, A. J. (1989). Ratings of managerial characteristics: Evaluation difficulty, co-worker agreement, and self-awareness. *Personnel Psychology, 42*, 235-261.

Mabe, P. A., & West, S. G. (1982). Validity of self-evaluation of ability: A review and meta-analysis. *Journal of Applied Psychology, 67*, 280-296.

Mann, F. C. (1965). Toward an understanding of the leadership role in formal organization. In R. Dubin, G. C. Homans, F. C. Mann & D. C. Miller (Eds.), *Leadership and productivity* (pp. 68-103). San Francisco: Chandler Publishing Company.

Marvin, M. A. (1995). *Survey research project* (Strategic Research Paper).

Mersmen, J. L., & Donaldson, S. I. (2000). Factors affecting the convergence of self-peer ratings on contextual and task performance. *Human Performance, 13,* 299-322.

Mumford, T. V., Campion, M. A., & Morgeson, F. P. (2007). The leadership skills strataplex: Leadership skill requirements across organizational levels. *The Leadership Quarterly, 18*, 154-166.

Mumford, M. D., Marks, M. A., Connelly, M. S., Zaccaro, S. J., & Reiter-Palmon, R. (2000). Development of leadership skills: Experience, timing, and growth. *The Leadership Quarterly, 11*, 87-114.

Nasby, W. (1989). Private self-consciousness, self-awareness, and the reliability of self-reports. *Journal of Personality and Social Psychology, 56*, 950-957.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

O'Connor, B. P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavior Research Methods, Instrumentation, and Computers, 32*, 396-402.

Owens, J. T. (1996). *A guide to the strategic leader development inventory.* Industrial College of the Air Force: National Defense University.

Podsakoff, P., & Organ, D. (1986). Self reports in organizational research: Problems and prospects. *Journal of Management, 12*, 531-544.

Schriesheim, C.A. & Eisenbach, R.J. (1995). An exploratory and confirmatory factor analytic investigation of item wording effects on the obtained factor structures of survey questionnaire measures. *Journal of Management, 21*, 1177-1193.

Society for Industrial/Organizational Psychology (2003). *Principles for the validation use of personnel selection procedures* (4th ed.). Bowling Green, OH.

Stogdill, R. M. (1974). *Handbook of leadership: A survey of the literature*. New York: Free Press.

Swiderski, M. J. (1987). Soft and conceptual skills: The often overlooked components of outdoor leadership. *The Bradford Papers Annual, 2*, 29-36.

Tabachnick, B. G., & Fidell, L. S. (2006). *Using multivariate statistics* (5th ed.). Boston, MA: Allyn and Bacon.

Thornton, G. C., III. (1980). Psychometric properties of self-appraisals of job performance. *Personnel Psychology, 33*, 263-271.

U.S. Department of the Army. (1999a). Army Leadership (Field Manual 22-100). Headquarters, Department of the Army, Washington, D.C.

U.S. Department of the Army. (2006). Army Leadership (Field Manual 6-22). Headquarters, Department of the Army, Washington, D.C.

U.S. Department of the Army. (2008b). Operations (Field Manual 3-0). Headquarters, Department of the Army, Washington, D.C.

U.S. Department of the Army. (1999b). U.S. Army vision statement. Available at: http://www.army.mil/vision/Documents/The%20-Army%20Vision.PDF.

U.S. Department of the Army (2008a). U.S. Army Training and Doctrine Command General TRADOC Vision. Available at http://www.tradoc.army.mil/about.htm

Vance, R. J., MacCallum, R. C., Coovert, M. D., & Hedge, J. W. (1988). Construct validity of multiple job performance measures using confirmatory factor analysis. *Journal of Applied Psychology, 73,* 74-80.

Van der Heijden, B.I.J.M.,& Nijhof, A.H.J. (2004). The value of subjectivity: Problems and prospects for 360-degree appraisal systems. *International Journal of Human Resource Management, 15* (3), 493-511.

Van Velsor, E., Taylor, S., & Leslie, J. (1993). An examination of the relationships among self-perception accuracy, self-awareness, gender and leader effectiveness. Human Resource Management, 32, 249-263.

Velicer, W. F., Eaton, C. A., & Fava, J. L. (2000). Construct explication through factor or component analysis: A review and evaluation of alternative procedures for determining the number of factors or components. In R. D. Goffin and E. Helmes, (EDs.), *Problems and solutions in human assessment*. Boston: Kluwer.

Woehr, D. J., Sheehan, M. K., & W. Bennett, Jr. (2005). Assessing measurement equivalence across rating sources: A multitrait-multirater approach. *Journal of Applied Psychology*, 592-600.

Wong, L., Bliese, P., & McGurk, D. (2003). Military leadership: A context specific review. *The Leadership Quarterly, 14*, 657-692.

Wong, L., Gerras, S., Kidd, W., Pricone, R., & Swengros, R. (2003). *Strategic leadership competencies.* Carlisle, Pennsylvania: Strategic Studies Institute of the U.S. Army War College.

Wright, P. L. (1996). *Managerial leadership*. New York: Routledge.

Wright, P. L., & Taylor, D. S. (1985). The implications of a skills approach to leadership. *Journal of Management Development, 4*, 11-18.

Yammarino, F. J., & Atwater, L. E. (1993). Understanding self-perception accuracy: Implications for human resource management. *Human Resource Management, 32*, 231-247.

Yukl, G. (2006). *Leadership in Organizations (6th Edition)*. Englewood Cliffs, NJ: Prentice-Hall.

Yukl, G., & Van Fleet, D. D. (1992). Theory and research on leadership in organizations. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (2nd ed.). Palo Alto, CA: Consulting Psychologists Press, Inc.

Zwick, W. R. & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin, 99*, 432-442.

Appendix A

Items from AZIMUTH

Communicating

1. **Does not provide clear direction.**
2. Explains own ideas so that they are easily understood.
3. Keeps others well informed.
4. Listens well.
5. Tells it like it is.
6. **Writes poorly.**

Decision-Making

1. **Delays decisions unnecessarily.**
2. Generates innovative solutions to unique problems.
3. **Ignores information that conflicts with own initial assumptions.**
4. Makes sound decisions in a timely manner.
5. Willing to revisit a decision when new information calls for it.

Motivating

1. Creates a supportive work environment.
2. Disciplines in a firm, fair, and consistent manner.
3. Inspires people to do their best.
4. Often acknowledges good performance of others.
5. Sets clear performance expectations.

Developing

1. **Does not encourage professional growth.**
2. Is an effective teacher.
3. Often uses counseling to provide performance feedback.
4. Provides opportunities to learn.
5. **Seldom delegates authority.**

Building

1. Actively participates in organizational/unit activities.
2. Encourages cooperation among team members.
3. Encourages organization unit/activities.
4. Focuses the organization/unit on mission accomplishment.
5. Treats others as valuable members.

Learning

1. **Becomes defensive when given critical feedback.**
2. Encourages open discussion to improve the organization/unit
3. Helps organization/unit adapt to changing circumstances.
4. Seems to be realistic about own personal limitations.
5. Willingly accepts new challenges.

Planning and Organizing

1. Anticipates how different plans will look when executed.
2. Develops effective plans to achieve organizational goals.
3. **Leaves key events to chance.**
4. Sets clear priorities.
5. **Unwilling to modify original plan when circumstances change.**

Executing

1. Completes assigned missions to standards.
2. **Does not meet mission timelines.**
3. Does Whatever is necessary (within ethical limits) to complete the mission.
4. Monitors execution of plans to identify problems.
5. Refines plans to exploit unforeseen opportunities.

Assessing

1. Accurately assesses the organization/Unit's strengths
2. Accurately assesses the organization/unit's weaknesses.
3. Makes organizational changes for no apparent reason.
4. **Rarely conducts after-action reviews.**
5. Takes time to find out what subordinates are doing.

Respect

1. Actively supports equal opportunity for all persons.
2. Creates a climate of fairness in the organization/unit.
3. **Excludes some from team activities.**
4. Treats others with respect.

Selfless Service

1. **Claims credit for others' work.**
2. Considers the needs of own and others' family members.
3. Places the welfare of the organization before own personal gain.
4. **Takes advantage of others to advance own career.**
5. **Takes privileges not allowed others.**

Integrity

1. **Behaves with questionable ethics.**
2. Demonstrates moral courage (does what is right).
3. **Is not sensitive to the ethical impacts of decisions.**
4. Is trustworthy.
5. Sets the proper ethical example for others.

Stability

1. Does not display extreme anger.
2. **Exhibits wide mood swings.**
3. Maintains calm disposition under stress.
4. Possesses an even temperament.
5. **Seems to behave unpredictably.**

Other

Demonstrates appropriate Soldier skills.
Is a clear thinker.
Is effective on the job.
Maintains effective interpersonal relations with others.
Physically fit for the job.
This person is a good leader.
This person is someone I would follow into combat.
The questions contained in the AZIMUTH were clear and understandable.
The questions contained in the AZIMUTH measure elements critical to excellence in leadership.
I am comfortable with the confidentiality of my answers using this procedure

Each item contains a seven-point scale ranging from “Extremely Good Description” to “Extremely Poor Description” with an additional option of “Have not observed”. For the classroom based version, ‘unit/organization’ was changed to ‘section.’ **Bolding denotes negative items that were reverse coded.**

Appendix B
Reliability of Original Leadership Dimensions

Table B-1

Intraclass Correlations for Ratings of Each Rating Source of Original Leadership Dimension Conceptualization

Dimension	Peer	Subordinate	Superior
Assessing			
(ICC 1, 1)	.29	.39	.51
(ICC 1, k)	.67	.76	.84
Building			
(ICC 1, 1)	.67	.83	.54
(ICC 1, k)	.91	.90	.86
Communicating			
(ICC 1, 1)	.35	.37	.43
(ICC 1, k)	.76	.78	.82
Decision-making			
(ICC 1, 1)	.31	.41	.58
(ICC 1, k)	.69	.77	.87
Developing			
(ICC 1, 1)	.43	.40	.35
(ICC 1, k)	.79	.77	.73
Executing			
(ICC 1, 1)	.46	.54	.40
(ICC 1, k)	.81	.86	.77
Integrity			
(ICC 1, 1)	.36	.64	.57
(ICC 1, k)	.74	.90	.87
Learning			
(ICC 1, 1)	.47	.49	.54
(ICC 1, k)	.81	.83	.86
Motivating			
(ICC 1, 1)	.75	.72	.71
(ICC 1, k)	.94	.93	.92
Planning and Organizing			
(ICC 1, 1)	.37	.49	.45
(ICC 1, k)	.75	.83	.80
Respect			
(ICC 1, 1)	.62	.43	.57
(ICC 1, k)	.87	.73	.84
Selfless Service			
(ICC 1, 1)	.43	.29	.59
(ICC 1, k)	.79	.68	.88
Stability			
(ICC 1, 1)	.49	.51	.65
(ICC 1, k)	.83	.84	.90

Appendix C

Required Raters of Original Leadership Dimensions

Table C-1

Number of Required Raters for Each Rating Source of Original Leadership Dimension Conceptualization

Dimension	Peer	Subordinate	Superior
Assessing	6	4	3
Building	2	1	2
Communicating	5	4	4
Decision-making	6	4	2
Developing	4	4	5
Executing	3	2	4
Integrity	5	2	2
Learning	3	3	2
Motivating	1	1	1
Planning and Organizing	4	3	3
Respect	2	3	2
Selfless Service	3	6	2
Stability	3	3	2
AVERAGE	6	3	3

Number of required raters was calculated by applying the Spearman-Brown Prophecy formula to the ICC (1, 1) values from Table B-1.